# 4. SAMPLING DESIGN

Sampling design is one of the major means by which the GRYN ensures scientific reliability and defensibility of our program. However, the details of the individual sampling designs are beyond the scope of this chapter, and are provided within individual monitoring protocols. Rather, this chapter will identify the major themes and concepts behind our sampling designs that have guided our choices for particular vital signs or protocols.

## THE PURPOSE OF A SAMPLING DESIGN

The NPS I&M Program provides information on the status and trends of our natural resources that is essential for the National Park Service to uphold its mission of preserving the national parks unimpaired for the enjoyment of future generations. The information used to determine the state of park resources must be made using reliable scientific information. Thus, the primary purpose of a sampling design is to ensure that the data collected are representative of the target populations and sufficient to draw defensible conclusions about the resources of interest (EPA 2002).

## BASIC DESIGN CONSIDERATIONS

### *General Sampling*

#### 1. PROBABILITY-BASED SAMPLING

Because a sample is used to draw valid conclusions about some larger population, it is imperative that the sample is representative of the population of interest (Lohr 1999). Three broad approaches to obtaining samples that are representative of the population include: probability-based sampling; judgment sampling; and convenience sampling. The GRYN considers probability sampling to be the most defensible because it applies sampling theory and some form of randomization in the selection of sample units (EPA 2002). This randomization ensures a reduction in potential bias from judgment or convenience sampling, thus increasing the validity of extending inference from a sample to the population of interest.

Common alternatives to probability-based sampling are judgment and convenience sampling. Judgment sampling employs expert knowledge in the selection of sampling units. Studies have shown that selection bias is common when judgment sampling is used (Edwards 1998, Stoddard et al. 1998, Olsen et al. 1999), although there remains some disagreement among ecologists and statisticians about the validity of using judgment sampling in some contexts (e.g., sentinel sites) (Edward 1998). Convenience sampling is generally based on factors such as ease of access and, thus, there is no assurance that samples collected in this manner will be representative of the target population. While convenience sampling is not considered a valid approach for the GRYN monitoring program, factors that improve efficiency of sampling (e.g., access) will be considered within the context of a probability-based sampling through stratification (see below).

#### 2. SAMPLING FRAME, SAMPLING UNITS

There are subtle differences in how some references define terms associated with sampling. There are probably even greater differences in how these terms are interpreted and applied on different projects. Figure 4.1 illustrates the use of these terms by the GRYN in the context of this report.

**Target Population—** The target population is a set of all of the units or elements for which inference is intended and should directly reflect the monitoring objectives.

**Sampling Frame**— The sample frame or sampling frame is a complete collection of the possible sample units (see below) from which the sample can be drawn. There are two types of sample frames commonly recognized: a list frame and an area frame. A list frame is a list of the potential sampling units along with their descriptive attributes. An area frame is typically designated by geographical boundaries within which the sampling

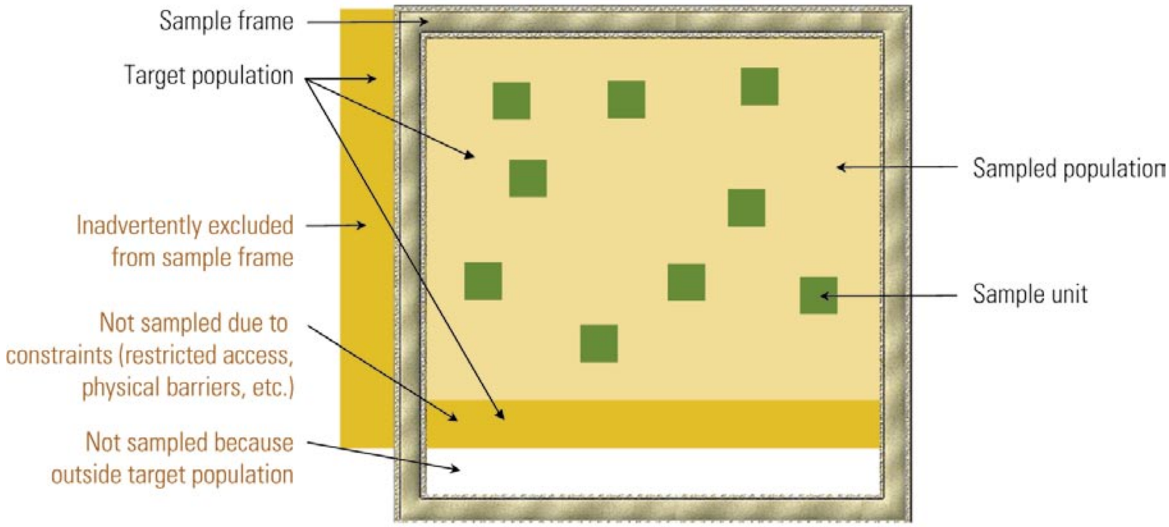


**FIGURE 4.1** Conceptual illustration of terms used to describe different units associated with entities being sampled (adapted from A.R. Olson [unpublished presentation] and Lohr 1999).

units are defined as subareas. Some designs (e.g., dual-frame designs, described below) use both frame types.

**Sampled Population—** The sampled population represents the actual population from which a given sample is drawn. As discussed below, ideally the sampled population would coincide with the target population and the sample frame, but a perfect overlap of these is rarely possible in environmental settings.

**Sample Units—** The sample units include all of the individual units contained within the frame that are actually sampled. Frequently, this concept appears to be more easily understood than it actually is under certain circumstances. For example, if the objective is to estimate the size of fish in a pond, then individual fish are the sample units. If, however, the objective is to estimate the proportion of native to exotic fish in a collection of ponds, then the sample units would be the ponds.

**Elements—** In some cases measurements may be taken on individual items within a sample unit. Thus, an element, sometimes referred to as an observational unit, consists of any item for which measurement is made or information is recorded (Schaeffer et al. 1990, Lohr 1999). These are typically individual plants or animals within a sample unit such as a transect, plot or grid cell.

**Note:** It is important to distinguish the sampling units from elements within a sampling unit because it is not uncommon for the number of elements to be incorrectly treated as if they represented independent replicate samples. This is a form of "pseudo replication" (Hurlbert 1984) and is a common source of statistical error in testing environmental effects.

Ideally, the sampled population and the sample frame would be equivalent to the target population for which inference is to be drawn. Unfortunately, numerous constraints exist that may preclude this from occurring (Figure 4.1), and, therefore, in some situations units within the sample frame and target population are not included in the sampled population.

*GRYN Example*

In the GRYN, constraints may result from safety concerns (e.g., bear closures), physical barriers or access limitations. It is also possible that part of the sample frame may inadvertently include units that are not within our target population. For example, our sample frame for monitoring whitebark pine is based on a map of whitebark pine generated from satellite imagery. If sites were erroneously classified as whitebark pine that actually did not contain whitebark pine trees, these would be included in our initial sample frame, but not sampled. It should be noted that when the sampled population does not coincide with the target population, that valid inference is limited to the sampled population.

## *Spatial Allocation of Samples*

There are a multitude of potential sampling designs for selecting a sample over space, although most are variations on a few basic themes. Following is a description of the major design themes and the specific variations on these themes will be discussed within individual monitoring protocols (see also Figure 4.2).

**Complete Census**— One special case of spatial sampling is a complete census, in which measurement is taken on all of the sample units within the population. As such, there is no sampling error that results from taking a sample (because all units are

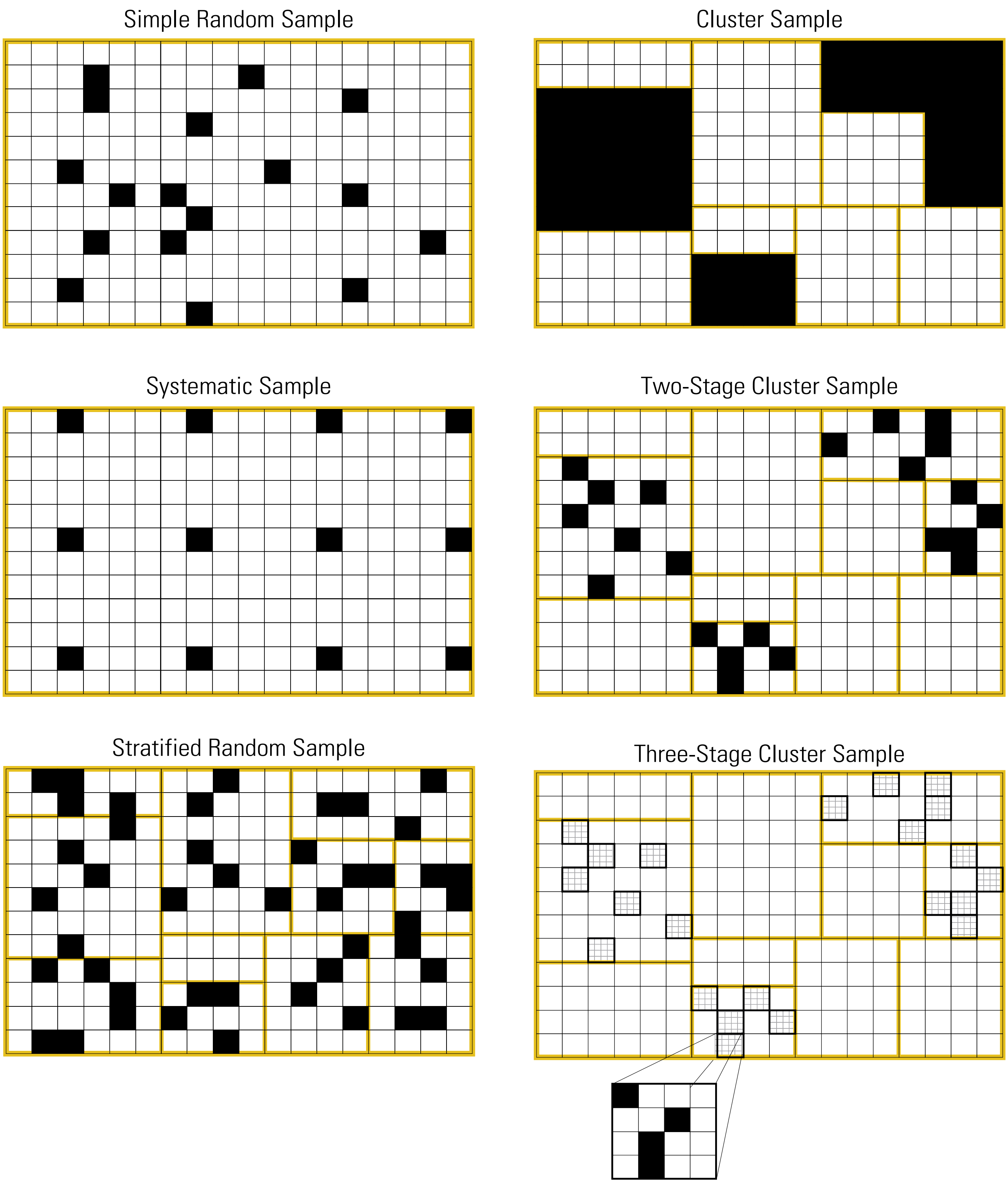


**FIGURE 4.2** Conceptual illustration of major spatial designs (adapted from Thompson [2002] and Lohr [1999]).

sampled) which is then applied to estimates for the entire population. However, a complete census may include measurement error associated with the measurement of each sample unit.

**Simple Random Sample—** In simple random sampling, (n) units are selected from a population of size (N) via a random process, such that every sample unit has the same probability of being included in the sample.

**Systematic Sample—** A sampling method in which one sample unit is typically selected at random and subsequent units are selected according to a systematic pattern. A common form of systematic sampling is randomly selecting one unit from the first k units in the sampling frame and every $k^{th}$ unit thereafter (Mendenhall et al. 1971).

**Stratified Random Sample—** In a stratified random sample the sampling frame is divided into mutually exclusive and exhaustive subpopulations called strata, from which n samples are randomly selected from each strata (Levy and Lemeshow 1999). There are several reasons for using stratified sampling design, including increased precision, increased efficiency and greater information about a particular subpopulation(s) (Cochran 1977, Lohr 1999). For increased precision, strata are typically selected such that the variation among units from the same strata is less than the variation among units from different strata (Thompson 2002). Increased efficiency may be based on such things as ease of access or administrative boundaries (Cochran 1977).

**TABLE 4.1** Advantages and disadvantages of major spatial design themes.

| Sampling Design | Major Advantages | Major Disadvantages |
|---|---|---|
| Complete Census | • No sampling error | • Seldom logistically or economically feasible<br>• Usually requires greater effort than is needed |
| Simple Random Sample | • Simple and straightforward analysis<br>• Doesn't require prior knowledge regarding sampling units. | • Can result in poor spatial distribution, particularly with small samples.<br>• Can be inefficient for rare or highly clumped resources |
| Systematic Sample | • Good spatial coverage<br>• Simple and straightforward<br>• Requires little or no prior knowledge regarding sampling units.<br>• Facilitates co-location of samples | • May not be as efficient as alternative designs if prior information about units is available.<br>• If properties of interest are aligned or there are periodicities with grid, then biased estimates are possible.<br>• A single systematic sample may not produce valid estimates of the standard error under some circumstances |
| Stratified Random Sample | • Can reduce costs and sample sizes<br>• Can increase precision | • Requires prior knowledge regarding sampling units.<br>• May reduce precision if criteria for strata assignment are uncorrelated. |
| Cluster Sample | • Can be cost efficient (i.e., it is often cheaper to sample all of the elements within a unit than to sample an equal number of elements at random)<br>• Can be feasible to construct a sampling frame, even when lists are difficult to obtain | • All of the elements within a cluster must be sampled<br>• Appropriate analyses are less straightforward.<br>• Lower precision than simple random or stratified sampling |
| Two-stage Cluster Sample | • Can be more efficient than single stage when clusters are too large or list units are homogeneous within clusters. | • Analyses more complex |
| Generalized Random-Tessellation Stratified (GRTS) | • Samples are spatially balanced<br>• Nested subsamples easily accomodated<br>• Good variance properties | • The underlying sampling process is less intuitive to understand than alternative sampling schemes.<br>• Software to use GRTS has only recently been made available to the public |

*GRYN Example*

In the GRYN, many resources are very difficult to access, which can greatly increase the cost and effort required for sampling. Stratification by access can be accomplished by treating areas within a different distance class from access points as different strata (e.g., close, moderate and far from access). Sites that are difficult to access can be sampled at lower frequencies but still be included within the sample. This enables a more efficient sampling effort and reduced cost without sacrificing the original scope of inference.

**Panel—** Refers to the group of sample units that are sampled during the same sample occasion (time block). For example, if sampling were conducted annually, then all of the units sampled in a given year would comprise the panel for that year. If all of the sample units were sampled every year, then there would only be a single panel for the design (Figure 4.3). During any given sampling occasion, either all of the sample units comprising a panel are sampled or none are sampled.

**Revisit design—** Refers to the plan or strategy for re-sampling panels over time.

**Cluster Sample—** Cluster sampling is an approach whereby selection is made of groups or clusters of units, called primary units, within which all of the secondary units are sampled (Levy and Lemeshow 1999). This approach is often used when it is difficult or impossible to enumerate all of the individual units within a sampling frame. Thus, enumeration of units is only necessary for the selected clusters.

**Multi-stage Cluster Sample—** Multi-stage cluster sampling is an extension of cluster sampling where a subset of the units within the primary units are sampled. For a two-stage cluster design, a sample of secondary units is selected, typically by a random process, from within the primary units. For a three-stage design, a sample of units is taken from the secondary units, and so on.

**Generalized Random-Tessellation Stratified (GRTS)—** The GRTS design uses a hierarchical randomization process to achieve spatial balance across regions and resources. GRTS samples also easily accommodate nested designs and allow units to be added efficiently after an initial sample has been drawn. Because GRTS samples achieve spatial balance without being evenly spaced, problems associated with correlations between systematic sampling and environmental gradients are reduced.

## *Temporal Allocation of Samples*

Following is a list of defined terms that pertain to temporal allocation of samples (terms have been adapted from McDonald 2003).

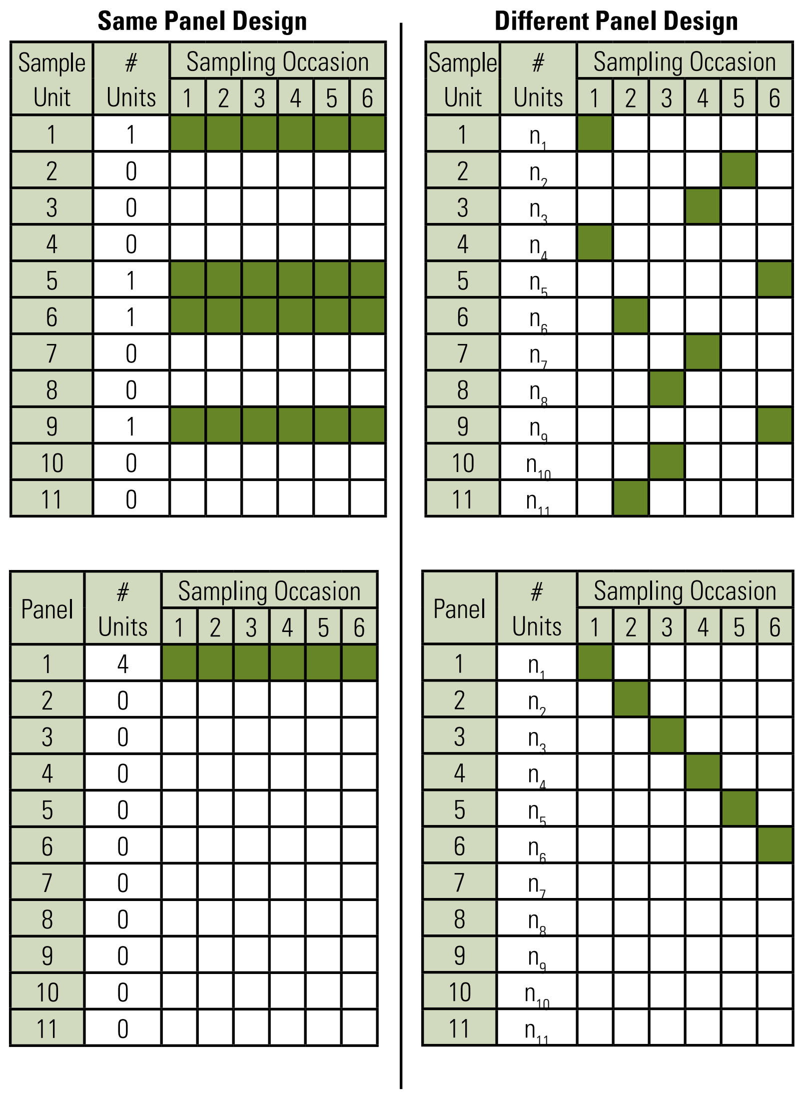


FIGURE 4.3 Graphical illustration of the relationship between spatial and temporal sampling. The Y axis of the upper diagrams represent revisit designs based on sample units, which can become quite cluttered when illustrating complex revisit designs. The lower diagrams represent revisit design-based panels, which are more efficient for illustration, but mask the spatial representation by condensing all of the sample units sampled at a given time into a single panel. To see this relationship, the number of sample units are shown for both graph types.

### I. MAJOR TEMPORAL (REVISIT) DESIGN THEMES

As with spatial designs, numerous temporal sampling (revisit) designs exist, with most being variations on a few basic themes. Following is a description of the major design themes for resampling over time that form the basis of our designs. These have also been illustrated in Figure 4.4. Specific variations on these themes for a given vital sign will be discussed within individual monitoring protocols.

**Complete Revisit Design—** Under this design, each sampling unit is revisited on each occasion (McDonald 2003). If the primary objective is to detect a linear trend over time, then this design is probably the most powerful (see discussion of power below) (Urquhart and Kincaid 1999). A primary disadvantage of this approach is that it is also probably the poorest for estimating the overall status, because the same sites are repeatedly visited rather than increasing the spatial representation by sampling sites at different locations. Other pieces of the design that must be taken into consideration include: whether or not repeatedly visiting a site can alter its response (e.g., habituation of animals, trampling, etc.); and replacement of units that are no longer usable (e.g., animals that have died, habitats that have changed types, etc.).

**Never Revisit Design—** Under this design, a different sampling unit is visited on a given sampling occasion and never visited again (McDonald 2003). Such designs are commonly used during inventories, where the primary objective is to estimate status. For that

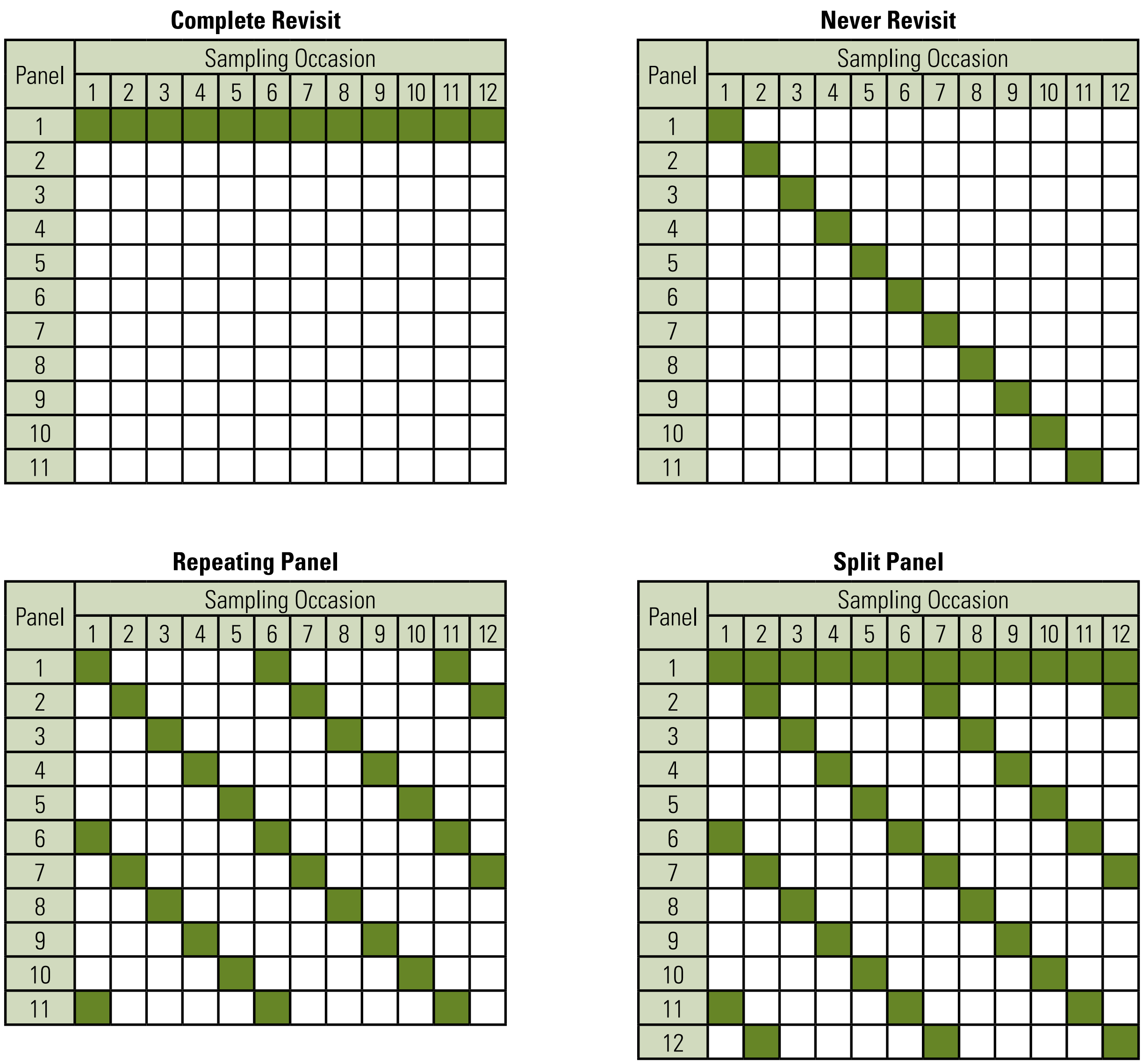


**FIGURE 4.4** Graphical representation of major revisit designs.

purpose, this design is efficient because it includes the greatest number of sites (Urquhart and Kincaid 1999). However, for monitoring change over time, it will likely be of limited value.

**Repeating Panel Designs—** These are designs in which a given survey panel is measured repeatedly over time. For the general case, the number of consecutive sampling occasions that a panel can be surveyed and the interval between consecutive samples can be varied and will depend on the specific monitoring objectives and time scales appropriate to meet those objectives.

**Split Panel Designs—** This is a design that partitions (splits) the panels into two or more revisit designs. This type of design enables different types of change to be detected (e.g., individual change and gross change). Such an approach also constitutes a compromise between emphasis on spatial and temporal variation. Typically, split panels entail an always visit design in combination with some other revisit design (e.g., repeating panel). The "always visit" design is the strongest for detecting temporal variation, but is weak for detecting spatial variation since the same panels are visited on each occasion. Combining this with an alternative panel can strengthen detection of spatial variation.

## *Sample Size Considerations*

### 1. MAGNITUDE OF CHANGE

Populations in the real world are dynamic, and change over time is to be expected. Even in the absence of some anthropogenic stressor it would be extremely unlikely for a given population to remain constant over time. Thus, to design a monitoring program whose primary purpose is to identify whether there has been a change over time can be rather trivial. What is more important is whether or not there has been a meaningful change (to the public and/or park managers), what has caused that change, and whether or not the resource is expected to change further.

To understand what constitutes meaningful change, it is essential to realize the difference between statistical significance and biological significance. Statistical significance relies on probability and is influenced by sample size. Thus, even minor changes (from a biological perspective) will be statistically significant if the sample size is large enough. Regardless of statistical significance, we would consider something biologically significant if it facilitates a major shift in the ecosystem structure or function such as a loss of one or more species, the addition of non-native species, changes in ecosystem processes, etc.

| | True | False |
|---|---|---|
| Reject $H_o$ | Type I Error ($\alpha$) | No Error |
| Do Not Reject $H_o$ | No Error | Type II Error ($\beta$) |

**FIGURE 4.5** Possible outcomes for null hypothesis test of no effect. Type I error occurs when the test indicates an effect (i.e., null hypothesis rejected) when there was no effect. Similarly, a Type II error occurs when an effect is not detected (failure to reject the null hypothesis) when there really was one.

From a monitoring standpoint, we are concerned with both statistical and biological significance in the sense that we want to know whether or not we are likely to detect a change statistically that we would consider biologically meaningful. To answer this, we need information about what level of statistical significance to we want to attain (i.e., our Type I error rate or $\alpha$) (see below), what level of change do we consider biologically meaningful that we hope to detect, and how variable is the resource that we are trying to estimate. With this information we can better determine the likelihood of detecting a change (statistically) that we would consider biologically meaningful.

### 2. TYPE I AND TYPE II ERROR, AND STATISTICAL POWER

In statistical terms, Type I and Type II error refer to erroneously rejecting (Type I error) or failing to reject (Type II error) a null hypothesis (Figure 4.5). With respect to monitoring, a Type I error indicates that there is a trend (the "null" hypothesis is that there is no trend) when none exists, while a type II error occurs when a trend is undetected. The "P value" (or $\alpha$ level) is the probability of making a type I error, while $\beta$ is a type II error rate. Statistical power refers to the probability of not making a type II error (or 1 - $\beta$). It is important to note that statistical power depends on what level of Type I error is acceptable, what level of change (i.e., departure from zero trend) you are trying to detect, and the relationship between variation of the resource you are measuring and the sample size used to detect the trend. Estimating power enables us to determine the sample size needed in order to detect a trend of a given magnitude with reasonable confidence.

## Sampling Rare Resources or Resources of Special Interest

Sampling rare resources is often problematic because most major design frameworks are inefficient for sampling rare resources. Increasing overall sample sizes to increase the likelihood that rare resources are included in a given sample can be effective but quite costly. Targeting specific resources in a list sampling frame can certainly improve the efficiency, but will greatly limit the scope of inference to those specific units that were targeted (i.e., the results cannot be generalized to the rare resource as a whole). Following are some design considerations that are considered for those vital signs that represent rare resources.

### 1. Stratified Sampling with Disproportional Allocation

One approach to ensuring that rare resource are adequately sampled is to partition the sample frame into strata such that one or more strata includes a high probability of containing the rare resource. This stratum can then be intentionally sampled with sufficient intensity so as to increase the likelihood of encountering the rare resource. This enables a more adequate characterization of the rare resource, but assumes prior knowledge about the distribution of the rare resource.

### 2. Dual Frame Designs

A dual frame design is one that incorporates more than one sample frame (Groves and Lepkowski 1985). A common approach to dual frame sampling is to combine a list frame and an area frame. The list frame contains known information about the targeted resources, such as known nest sites, specific geothermal features, etc. Sampling these known units can provide valuable information about changes in those specific resources over time, but will not allow inferences to be generalized to the rare resource as a whole. Yet by adding in an area frame, a probabilistic design (that would be inefficient on its own) is then combined with the list frame allowing some inferences to be extended beyond the specific targeted resources (Haines and Pollock 1998).

*GRYN Example*

One of the specific monitoring objectives identified for the amphibian vital sign was to monitor changes in occupancy of boreal toad breeding sites. Boreal toads are quite rare in the GRYN, and existing data indicate that a probabilistic sample to monitor potential changes would be insufficient to estimate changes in the primary parameter of interest (proportion of catchments occupied). Given that there are a limited number of known breeding sites for this species, meaningful estimates of change over time necessitate targeting known sites. While this approach should provide reasonable estimates of the change in occupancy of these sites over time, a disadvantage is that the inference would be limited to these sites. However, for other objectives, a cluster sampling design of an area sampling frame will be used for some of the other amphibian objectives. Including both frames in the design will allow for inference about occupancy of toads at sites that are currently unknown and, thus, not part of our list frame.

### 3. Adaptive Sampling

In most traditional sampling designs, the selection of sampling units is not influenced by what is observed during the sampling. In contrast, adaptive sampling entails the selection of units that may be influenced by the value or type of unit selected (Thompson and Seber 1996). Typically, a decision rule is established a priori that triggers a change in the sampling as it occurs. (Figure 4.6). Thus, adaptive sampling can be an effective design for rare resources, particularly if prior information about the distribution of that resource is poorly known.

Adaptive sampling can be incorporated into a wide variety of traditional designs (e.g., simple random samples, systematic samples or cluster samples). However, it can also introduce bias, which needs to be accounted for with estimators developed for adaptive designs (Thompson 2002). Although none of the vital signs currently have adaptive sampling included in their protocols, it remains a possible sampling design that could be used.

# Integration

## The Need for Integration

### 1. Integration Across Networks

The I&M Program was intended from the outset to focus on information needed by park managers for understanding and managing our network parks. However, it was also intended from the outset that some subset of the selected vital signs would provide information at scales larger then the GRYN (e.g., water quality). Thus, an additional design consideration has been whether or not there is a need, value or expectation for implementing designs that can be scaled up to levels beyond the GRYN.

*GRYN Example*

Water quality is an example intended from the outset to provide information for local park managers, while at the same time providing infor-

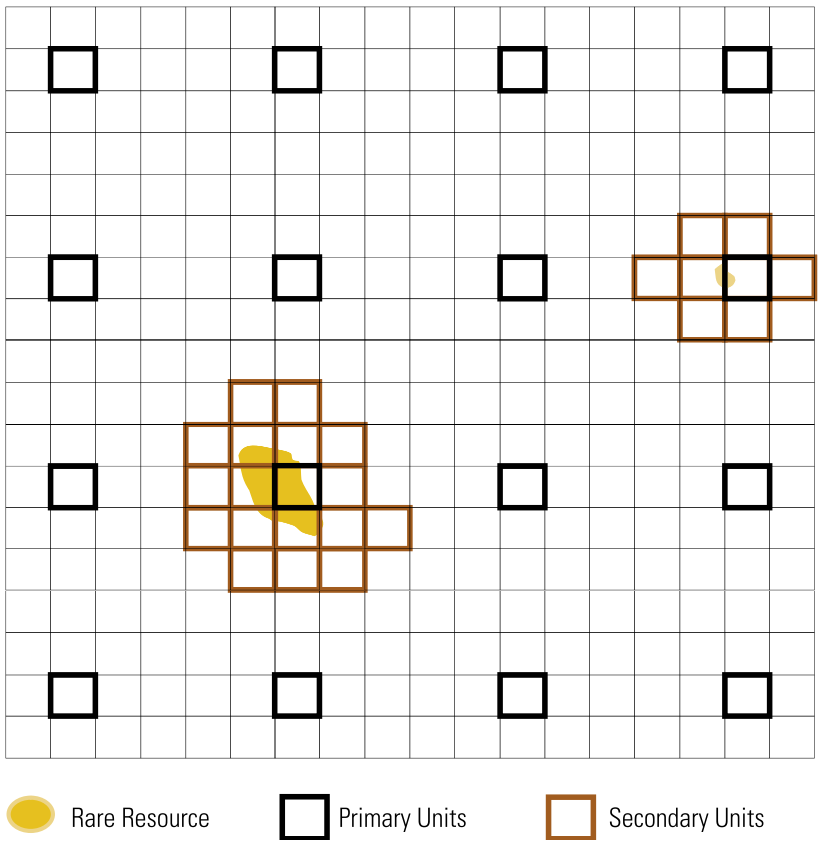


**FIGURE 4.6** Conceptual illustration of adaptive sampling where the primary units are drawn from a systematic sample, and the secondary units are drawn when a rare resource is encountered, and continue until such time that the resource is no longer encountered.

mation on the quality of waters within national parks at larger scales (e.g., networks, regions and national). To facilitate this goal, the GRYN has entered into a partnership with the Rocky Mountain Network to develop jointly a sampling design that will enable interferences to be scaled up or down from local to more regional scales (see section below on "A Generalized Design for Aquatic Resources").

### 2. INTEGRATION ACROSS AGENCIES

Although the inventory and monitoring program is a National Park Service endeavor, many of the vital signs that we will monitor cross over jurisdictional boundaries, and concerns about these vital signs are often shared by other agencies. There can also be increased efficiency and broader application through cooperative efforts among agencies. Thus for vital signs that have a common interest among agencies and organization, we will attempt to coordinate, and where possible, collaborate, with other agencies for a more effective monitoring program.

*GRYN Example*

Concerns about whitebark pine have been shared by several agencies within the Greater Yellowstone Ecosystem (GYE) for several years. Several efforts to assess the status of white pine blister rust in whitebark pine communities have been conducted in the GYE, although efforts to monitor whitebark have generally been sporadic with little coordination among agencies. Consequently, the GRYN has joined forces with several organizations including the Forest Service (Forest Health Monitoring Program and six National Forests), the USGS (Northern Rocky Mountain Science Center), the Interagency Grizzly Bear Study Team (a cooperation among USGS, NPS, USFS, and the states of Wyoming, Idaho and Montana), and the Greater Yellowstone Coordinating Committee (a cooperation among NPS, USFS and USFWS). In this effort, we have formed a small working group that meets on a regular basis. This group is made up of representatives from the cooperating agencies. The group provides a forum for discussion and resolution of ideas ranging from agreement on the specific monitoring objectives to development of a cooperative design. The result will be to implement an interagency whitebark pine monitoring effort for the Greater Yellowstone Ecosystem.

### 3. INTEGRATION AMONG VITAL SIGNS

Vital signs are not environmentally and ecologically independent entities. Rather, they are often the products of complex interactions among other vital signs and/or other ecosystem components or attributes. Without some consideration of how our vital signs interact, the GRYN program has no added value apart from the sum of its parts. Thus, consideration is needed as to how the parts fit together as a whole. Some level of integration among vital signs is needed if we expect to (1) understand the dynamic responses to changes in drivers or stressors, (2) understand the interaction effects among vital signs, and (3) reduce the confounding effects of other vital signs in the interpretation of a given vital sign.

## Considerations for Use of a Generalized Overall Design

One solution for achieving some level of integration is a generalized overall design (i.e., one used for several vital signs). A systematic design is probably the most reasonable for common use among vital signs. Using simple random sampling or cluster sampling will present the problem of which units to select. A reasonable unit (sample unit or cluster) for one vital sign may not be reasonable for another. In contrast, a systematic design would enable a distribution across space regardless of the units. Systematic designs are typically relatively simple and robust; they have reasonable precision; and they can be an effective way to ensure that areas are sampled in proportion to their size. Use of a common design among networks can enable scaling up or down from local to more regional inferences. Co-location of samples from different vital signs within a common design can enable increased ability for assessing the effects of drivers or stressors, as well as interaction effects (see below). However, adopting a generalized overall design is not without its limitations. Depending on how much overlap there is among vital signs in space and time, a generalized overall design may be inefficient compared to some alternative directed at sampling a particular resource. There are other considerations, such as compatibility with existing efforts or with partner organizations. A summary is presented below of the primary factors contributing to a decision about whether or not to incorporate a given vital sign of the GRYN within a generalized overall design, or whether an alternative approach would be warranted.

### 1. National vs. Local Objectives

Virtually any reference regarding environmental and ecological monitoring will emphasize the importance of understanding the objectives of monitoring when developing a sampling design (e.g., Hellawell 1991, Spellerberg 1991, Olsen et al. 1999, Noon 2003). The value of an overall systematic design is probably at its greatest in large-scale monitoring programs (e.g., Forest Inventory and Analysis, Forest Health Monitoring, Environmental Monitoring and Assessment Program, Global Observation Research Initiative in Alpine Environments, etc.) whose goals are heavily focused on being able to scale up inferences from local to more regional or global scales. The NPS I&M Program differs from these other large-scale monitoring programs by emphasizing the information needs of individual parks that are linked together into networks. While maintaining the ability to detect regional and national level trends is desirable, the first priority of the GRYN is to meet the local information needs of the parks.

### 2. The Role of History and Existing Efforts

Many of the vital signs selected by the GRYN have also been selected for monitoring as part of cooperative efforts with other organizations. Consequently, the GRYN must consider the advantages and disadvantages of being part of a cooperative or pre-existing sampling design, as opposed to fitting within a generalized GRYN design.

*GRYN Example*

The Amphibian Research and Monitoring Initiative (ARMI) began in 2000 as a national program (coordinated by the USGS), whose aim is to better understand the dynamics of amphibian population trends. ARMI is currently establishing a north/south transect along the Rocky Mountains that would include Glacier, Yellowstone, Grand Teton and Rocky Mountain National Parks, among other locations. Given that the ARMI design is based on watershed units, rather than a systematic grid, our options are to design our amphibian monitoring program to be consistent with the Rocky Mountain Transect of ARMI, or to deviate from this approach to be part of a more generalized design for the GRYN. Given that our objectives for amphibians are consistent with those of ARMI, we believe that our best option is to conform to the ongoing effort rather than to deviate from the design currently in place.

### 3. Efficiency of Sampling Rare Resources

Another reason that we may chose not to adopt a generalized systematic approach is efficiency for some vital signs. For example, rare resources can be poorly represented in systematic designs. Encountering rare resources can be accommodated to some extent by increasing the overall sample size to better ensure that rare resources are included. However, such a solution can be costly in terms of effort and money, and it doesn't ensure adequate sampling of rare resources. The problem of sampling rare resources in a simple random or systematic design can be offset by stratification, although this assumes both that sufficient information exists to enable effective stratification and that there are not other resources concurrently being sampled that would not be conducive to a particular stratification for the rare resource.

### 4. Feasibility and Need for Co-location of Samples

One consideration in the sampling design(s), regardless of whether or not a generalized design is used, is whether or not samples for different vital signs should be physically co-located. Co-location of samples can facilitate assessment of the response to drivers or stressors and interaction effects. Under some circumstances co-location can also aid in the interpretation of confounding effects and increase efficiency of sampling. However, co-location of samples is not a panacea for ecological insights, and the costs and benefits need to be considered. To decide whether or not samples warrant co-location the GRYN considers: (1) the specific objectives of the vital sign(s) being sampled, (2) the feasibility of co-locating samples, (3) the probability of expected increased insights, and (4) the compatibility of domains and scales (see below).

One tool used by the GRYN to assess possible co-location of vital signs is conceptual models that focus on associations among the vital signs. For example, the simple conceptual model of whitebark pine presented in Chapter 2 (Figure 2.2) reveals several potential linkages among GRYN vital signs. Forest insects and disease (e.g., mountain pine beetle and white pine blister rust) and fire may have important influences on whitebark pine. Similarly, large carnivores (i.e., grizzly bears) and landbirds (i.e., Clark's Nutcrackers) also have strong associations with whitebark pine. Grizzly bears feed extensively on whitebark pine seeds. Clark's Nutcrackers also feed extensively on whitebark pine seeds and also play a major role in seed dispersal of whitebark pine. Thus, the feasibility and benefits of co-locating samples of these other vital signs and whitebark pine should be considered.

Another tool used to assess whether or not sampling from different vital signs should be co-located within a generalized design is a table of the overlapping domains (Table 4.2). This table summarizes some of the factors influencing feasibility of co-locating samples, including (1) geographic extent [i.e., parks], (2) aquatic vs. terrestrial system (3) primary habitat type and (4) potential for major partners or collaborators that may require design constraints. From such a table it can be seen that there is considerable overlap in some vital signs across these domains. For example, several of the aquatic and water quality vital signs have substantial overlap across all of these domains. Such overlap would indicate high feasibility for co-locating samples within a generalized design. Other vital signs have little overlap with others in these domains. For example, in the context of the GRYN, soil structure and stability is primarily focused on biological crust soils in aridland habitats of BICA, which may reduce its feasibility for inclusion within a generalized design for the entire network.

## GREATER YELLOWSTONE NETWORK DESIGNS

Based on the considerations described above, it was determined that a single overall design was not warranted for several GRYN vital signs, with the exception of aquatic resources (described below). Here we describe the general designs for those vital signs for which the development has reached the design stage.  We have grouped these according to the major spatial design themes.

**TABLE 4.2** Domains of each vital sign currently under development that are used to assess the feasibility of co-locating samples.

| Vital Signs | Park | | | Aquatic vs Terrestrial Resource | Habitats | Major Collaboration with other organizations |
|---|---|---|---|---|---|---|
| | BICA | GRTE | YELL | | Zone | |
| Climate | X | X | X | Terrestrial | Multiple | Ongoing |
| Soil structure and stability | X | | | Terrestrial | Aridland | |
| Arid seeps and springs | X | | | Aquatic | Wetlands | |
| Steamflow | X | X | X | Aquatic* | Perennial lakes/streams* | |
| Water chemistry | X | X | X | Aquatic* | Perennial lakes/streams* | |
| Aquatic invertebrate assemblages | X | X | X | Aquatic* | Perennial lakes/streams* | |
| Invasive plants | X | X | X | Terrestrial | Multiple | |
| Exotic aquatic assemblages | X | X | X | Aquatic* | Perennial lakes/streams* | |
| Whitebark pine | | X | X | Terrestrial | Sub-Alpine | Ongoing |
| Amphibians | X | X | X | Aquatic | Wetlands | Ongoing |
| Landbirds | X | X | X | Terrestrial | Multiple | Ongoing |
| Land use | X | X | X | Terrestrial | Disturbed/Developed | |

* Shared domain indicated in gold.

## Complete Census

### 1. Land Use

The primary sampling units for this vital sign are counties (agricultural) and 1 m2 township/ range/section (TRS), for which information is recorded annually for each home located outside of city boundaries. In each case all of the sampling units within the GRYN will be measured, thus comprising a census of the units, rather than a sample of the units. However, this does preclude measurement error that arises when the responsible agency measures the parameter of interest within each unit.

The temporal design for monitoring land use will be a repeating panel design, with each panel sampled during one sampling occasion (year) and then sampled again after five years (housing density and agriculture) or ten years (roads). This temporal design is based on the anticipated rates of change for this vital sign, as well as the intervals of measurement by the responsible agencies.

## Systematic Designs

### 1. Landbirds

The general sampling design for landbird pilot effort at GRTE is a two-stage systematic design, where a systematic grid was overlaid on a GIS sample frame of the targeted habitat types with a random start point. The grid was scaled to enable the approximate number of transects (with an oversampling rate of approximately 25%) to occur over the entire area. This enabled a reasonable degree of spatial balance. Grid points were randomly selected from this overlay and the selected points served as the starting point of a 2 km transect with a random vector from the start point. Distance sampling (either point or line transects) was then used to sample each secondary unit (transect). An alternative to this design would have been a GRTS design, although the software was not generally available at the time selection was made. A GRTS design may be used if the pilot transects are not used for the final transects or when the pilot effort is extended to full implementation.

The temporal design for landbird monitoring is an always visit design. The abundance of landbirds at a given site can be highly variable from year to year. Thus, revisit designs that have intervals between sampling occasions (years) can provide spurious results if they happen to fall on occasions that the parameter of interest is particularly high, low or coincidentally different. Sampling at intervals greater than one year can also greatly reduce our ability to interpret results in light of a fluctuating environment

## Cluster Designs

### 1. Amphibian Monitoring

Amphibian monitoring will be conducted in collaboration with the USGS Amphibian Research and Monitoring Initiative (ARMI). Within ARMI, a unified effort for monitoring amphibians along the Rocky Mountains from Colorado to Montana has been developed (Corn et al. 2005). The general design will be a single stage cluster design, with unequal probability of samples. The primary sampling units are hydrologic catchments equivalent to what would be approximately an 8th-order hydrologic unit code (HUC). Within each primary unit all wetlands will be surveyed. Identification of all wetlands requires an extensive ground search in addition to remote-sensing applications. Consequently, it is a prohibitively laborious task. Using cluster sampling only requires that this task be accomplished for the selected clusters. The size of the primary units that are currently being used resulted from an extensive collaborative effort with the USGS EROS Data Center as well as field testing as part of a pilot effort. The size of these units was intended to achieve a balance between having a sufficient number of wetlands to be efficient in detecting the presence of amphibians, but to eliminate oversampling of wetlands so that variation in occupancy would be difficult to detect and sampling could not be efficiently accomplished by a small crew within a short period of time (1-3 days). The general suitability of hydrologic units based on the quantity of NWI wetland types within each unit will be used to define unequal sampling probabilities. This is necessary because most hydrologic units within the parks have low value for amphibians. Thus, using an unequal inclusion probability will enable us to invest most of our resources to those units most important to amphibians.

The temporal design for amphibian monitoring will be an always revisit design. This temporal design was based on a collaborative decision with the USGS ARMI Program that estimating change over time within sampling units will be most informative, provided that we have a reasonable spatial representation in the initial panel. Preliminary attempts at defining clusters have indicated that this condition would be satisfied, given that the areal size of the units is carefully chosen (see above).

### 2. Whitebark Pine Monitoring

An existing protocol has been developed by the Whitebark Pine Ecosystem Foundation (Tomback et al. 2004), although modification was needed to meet GRYN objectives and I&M standards, particularly related to site selection. We have been working with partner organizations (USGS, U.S. Forest Service, the Greater Yellowstone Coordinating Committee, and the Statistics Department of Montana State University) to make revisions that will meet NPS standards, yet will still make use of those parts of the existing protocol that are acceptable. The resulting design from this effort is a two-stage cluster design with stands (polygons) of whitebark pine comprising the primary units and 10x50 m plots being the secondary units. The primary units (forest stands of whitebark cover classes) are selected from a sample frame derived from a GIS layer of predicted whitebark pine distribution. Within these stands, secondary units (plots) are selected from random points such that these plots comprise a subset of the potential plots within any given stand. Initially, we are sampling two secondary units within each stand to determine the extent of within-stand and between-stand variations. This may be modified at a later date if our results indicate that within-stand variation is such that a more efficient scheme would be to sample a greater number of stands. Within the secondary units, all live whitebark pine trees >1.4 m height within the transect are individually marked for future revisits to determine change in status of blister rust infection and survival.

The temporal design for monitoring whitebark pine will be a repeating panel design, with each panel sampled during one sampling occasion (year) and then sampled again after several years (probably five). This temporal design works well for this vital sign because white pine blister rust (the focus of our monitoring objective) is a slow acting pathogen that has relatively little inter-annual variation. Thus, sampling a given panel every year would be extremely inefficient. Sampling panels at longer intervals allows us to develop a sufficient sample size over several years while maintaining a reasonable ability for potential changes to be detected.

## *Generalized Random-Tessellation Stratified (GRTS) Design*

### 1. WATER CHEMISTRY

The details of our GRTS design for water chemistry are discussed in greater detail in the following section on "A Generalized Design for Aquatic Resources", but the general framework will be a dual-frame design where a targeted list of fixed sites is used in combination with probabilistic sampling using a GRTS design.

The temporal design will be a split-panel design where an always visit design will be used in combination with a repeating panel, such that the sites sampled every year will help to interpret the temporal variation that may be confounded with spatial variation from a repeating panel. For YELL, the repeating panel may be based on hydrologic basins because of logistical constraints of access and permits. These constraints are less prevalent at the other units where the rotating panels are not restricted to specific basins.

### 2. STREAMFLOW

A field measurement of discharge will be made at all sample sites of "flowing waters" within the same design framework as water chemistry.

### 3. AQUATIC INVERTEBRATE ASSEMBLAGES

A field measurement of aquatic invertebrate assemblages will be made at the same sample sites as water chemistry using the same general design framework.

### 4. EXOTIC AQUATIC ASSEMBLAGES

We anticipate sampling for exotic aquatic assemblages at either all or a subset of the sites sampled for water chemistry using the same general design framework. One of the advantages of GRTS is that a subsample that maintains the spatial balance of the overall sample can be collected.

## *Non-Probability Sampling*

### 3. CLIMATE

Unlike most vital signs, GRYN climate has been monitored continuously for over 100 yr. There is also a legacy network of monitoring stations maintained by a variety of state and federal agencies. Most of the existing sites were selected using a professional judgment selection process. Selection of sites through a probability sample in this case would not be practical given access and other logistic constraints. Further, changing the existing sites at this point could severely compromise the existing legacy of climate data for the GRYN. Consequently, protocol development and design for this vital sign is focusing on evaluating the following: (1) if the legacy network provides adequate sampling of spatial and temporal variability in GRYN climate and (2) how best to address shortfalls in the current system.

Our basic approach involves a detailed analysis of existing climate monitoring stations in the GYE to determine if:

1. Current stations in the GRYN can adequately capture the key spatial and temporal components of climate variability in the region.
2. Strata of management interest or scientific importance are being adequately sampled.

## A GENERALIZED DESIGN FOR AQUATIC RESOURCES

Based on the considerations described above, it was determined that a generalized overall design was warranted for several aquatic resources within the GRYN. A summary of such a design is presented below.

### *General Overview*

Several considerations influenced our choice of a specific generalized design. These included:

**Inferences scalable from individual parks to inter-network—** The final design must be able to accommodate inferences at a local scale (e.g., parks) as well as at more regional scales.

**Good spatial representation and dispersion—** Having a good spatial representation of targeted water bodies will help ensure the inferences to parks or higher are reliable. Similarly, water bodies that are connected or even in proximity to each other are subject to the same environmental influences and are not likely to be independent. Thus, a good spatial representation includes having reasonable dispersion of our sample.

**Complete and accurate sample frame—** To be consistent with other aquatic programs, the national hydrologic database (NHD) (USGS 1999) is the preferred sample frame for the GRYN. However, the NHD does not always provide accurately identified perennial streams, currently targeted by the GRYN objectives (Stevens and Olsen 2004). Thus the chosen design needed to be able to account for this problem.

**Complete availability of all potential sampling units within the sample frame—** Another potential problem to be considered is that for any reasonable sample frame, there are likely units that are unavailable within the GRYN because of safety constraints (e.g., bear closures or avalanche danger) or resource protection (e.g., nesting areas for sensitive species). Thus, the chosen design needs to be able to accommodate this problem.

Systematic designs are generally well suited for obtaining good spatial representation (Cochran 1977). Systematic sampling can also generally perform well in the presence of spatial autocorrelation, although under some circumstances a stratified random sample may be superior (Cochran 1977). In general, systematic sampling should also be well suited to scaling of inference from local to more regional scales. However, incomplete or inaccurate sample frame and the unavailability of some sample units can be problematic for some systematic designs. Stevens and Olsen 2004 present a systematic design called the Generalized Random-Tessellation Stratified Design (GRTS) which is well suited to accommodate these concerns about the sample frame. This design was also developed in the context of, and has been applied to, water quality monitoring; thus solutions for many potential problems that might arise have already developed.

### *Properties of the Generalized Random-Tessellation Stratified (GRTS) Design*

Details of the GRTS Design and how it works to achieve spatial balance are beyond the scope of this report, but have been reported in Stevens and Olsen (2004). Essentially, the GRTS design uses a hierarchical randomization process to achieve spatial balance across the region and the resource (Figure 4.7). A sample frame is created; in this case the NHD. A grid is randomly overlaid on the frame and

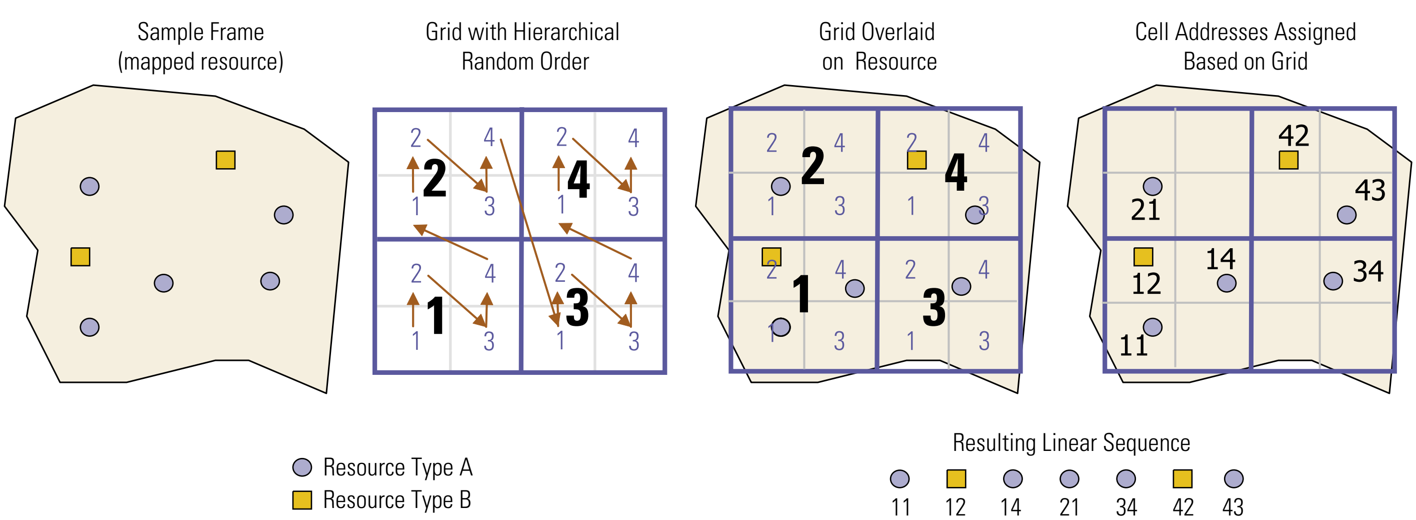


**FIGURE 4.7** Graphical representation of the steps leading up to selection of sample units using the generalized random-tessellation stratified design (adapted from Stevens and Olsen 2004, unpublished presentations).

subdivided until there is only one sample unit per cell. Cell addresses are assigned via a hierarchical random process, and each sample unit is assigned to its corresponding cell address, creating a linear sequence of sample unit cell addresses. By reversing the order of address digits and re-sorting this sequence, a systematic sample can be drawn with a random start point that maintains the spatial balance of the sample. Some of the resulting properties of this design that make it an attractive choice for water quality monitoring are:

- Sample is spatially balanced across the resource, resulting in improved precision and a more 'realistic' suite of statistics
- Spatial balance is maintained even at different sampling intensities and among samples and subsamples
- Nested subsamples are easily accommodated, which facilitates different suites of indicators to be measured at different subsets of sample sites. This can be important for accomplishing multiple objectives within the same general design
- Enables design-based estimators and their variances
- Applicable to point, network or areal resources
- Stratification and unequal sampling probabilities of subpopulations are easily accommodated. This can improve precision of estimates as well as increase the efficiency of sampling.

## *The Probability of Selecting Sample Units within the GRYN*

One approach to probability sampling is to assign equal weight to all sample units such that any particular unit has the same probability of being selected. However, there are many reasons why this may not always be the best approach. Taking into account groups of sample units can allow for inferences to be made for the group, increase precision of estimates, increase the efficiency of sampling, etc. There are two primary means by which the GRYN would consider groups, stratification and subpopulations. Stratification treats each group as a separate population for which samples are drawn independently for each stratum. Likewise, inferences are made for each stratum, but estimates can be combined across strata and weighted appropriately. An alternative to stratification is to designate groups of interest as subpopulations, which may have different probabilities of inclusion in the sample for each subpopulation. This approach can be an effective way to achieve some of the benefits of grouping (e.g., improved precision or efficiency), particularly when specific inference for the group is not essential. Tentative groups being considered for each of these approaches are listed below.

### 1. STRATIFICATION

Because parks are a basic unit of management, having estimates of water quality at the park level is essential. To ensure that a sufficient sample is obtained within each park to enable inference at that scale, the GRYN will stratify water quality sampling by parks.

### 2. SUBPOPULATIONS

The final details of what subpopulations will be recognized are pending further discussion, but tentative subpopulations are:

**Access Class—** Many locations within the GRYN are extremely difficult or costly to access, and considering access class as a subpopulation may be one solution to this difficulty. Units far from roads, trails and overnight facilities could be assigned sampling weights smaller than more accessible units.

**Major Watershed—** It is likely that major watersheds will constitute subpopulations. In the GRYN, features such as geothermal activity tend to correspond with particular watersheds.

**Strahler Order—** Similarly, the stream order (e.g., Strahler) has also been found in many systems to correlate well with important basin properties and is a likely candidate for consideration as a subpopulation unit.

**Perennial/Non-perennial Streams—** If there is a decision to include non-perennial streams, then this would likely constitute a subpopulation.

## *A Dual Frame Component*

As previously discussed, existing cooperative efforts need to be taken into consideration. In the case of the water resources vitals signs considered under the GRTS design, there are existing data from fixed stations that function in the context of integrator sites of NAQWA (Shelton 1994). Because the locations of these sites were specifically selected, inferences from such sites cannot reliably be extrapolated to the entire parks, the GRYN, etc. However, inferences about changes over time at these sites are quite legitimate, even if the spatial extrapolation to other sites is not reliable. Such continuous records over time are also quite valuable for other NPS programs (e.g., monitoring geothermal activity). Thus, there would be a great loss of valuable long-term information by abandoning these sites, even if a GRTS design is adopted from more generalized inference. A dual-frame approach will be considered that enables stronger inferences about temporal changes from existing fixed sites to complement broader scoped inferences from probabilistic sampling via the GRTS design.

## RESEARCH VS. MONITORING

The distinction between monitoring and research is not always clear. Monitoring is generally focused on the detection of changes or trends, whereas ecological research is focused more on the causes or associations of ecological patterns or processes. Monitoring is typically carried out over long time frames; whereas research is typically, but not exclusively, more limited in duration. Research can be conducted over relatively long time scales (e.g., the Long-term Ecological Research [LTER] Program), although it is still focused on answering questions, rather than estimating status and trends. In its "purest" form, research incorporates controlled experiments with random assignment of experimental treatments (Figure 4.8). In contrast, monitoring has typically entailed descriptive surveys of status and trends, occasionally including correlation surveys, and sometimes quasi-experiments, which can be considered as hypothesis generating, rather than hypothesis testing.

One goal of the vital signs monitoring program is to "monitor park ecosystems to better understand their dynamic nature and condition and to provide reference points for comparisons with other, altered environment" (National Park Service 2004). Accomplishing an understanding of the dynamics of park ecosystems is not likely to be accomplished by only estimating status and trends. This presents somewhat of a paradox in that monitoring alone may not be able to effectively achieve one of its primary goals (understanding of ecosystem dynamics), while at the same time, use of the I&M program for research purposes is neither practical nor its intended purpose. At the very least, the vital signs monitoring program is intended to work in conjunction with research to gain this understanding. Many of our vital signs were selected primarily because they are a major driver or stressor on park resources. With this in mind, we have described below a few ways that the GRYN program has considered the complementary roles of research and monitoring in the design of our program.

### *Identification of Research Needs*

As specific monitoring objectives for each vital sign are developed, it is determined whether the objective is better suited a research objective, rather than part of the I&M program. In some cases, research may even be needed to facilitate the formulation of meaningful monitoring objectives. This provides a source for proposing and/or prioritizing research conducted by or for the parks.

### *Confounding Variation and Auxiliary Variables*

Many vitals signs were selected because they are a known or suspected agent of change (e.g., ecosystem driver or stressor). In a research context, such variables are frequently incorporated as explanatory variables with the intent to determine if there is an association with a response variable of interest. However, there are additional reasons to include auxiliary variables that may not be vital

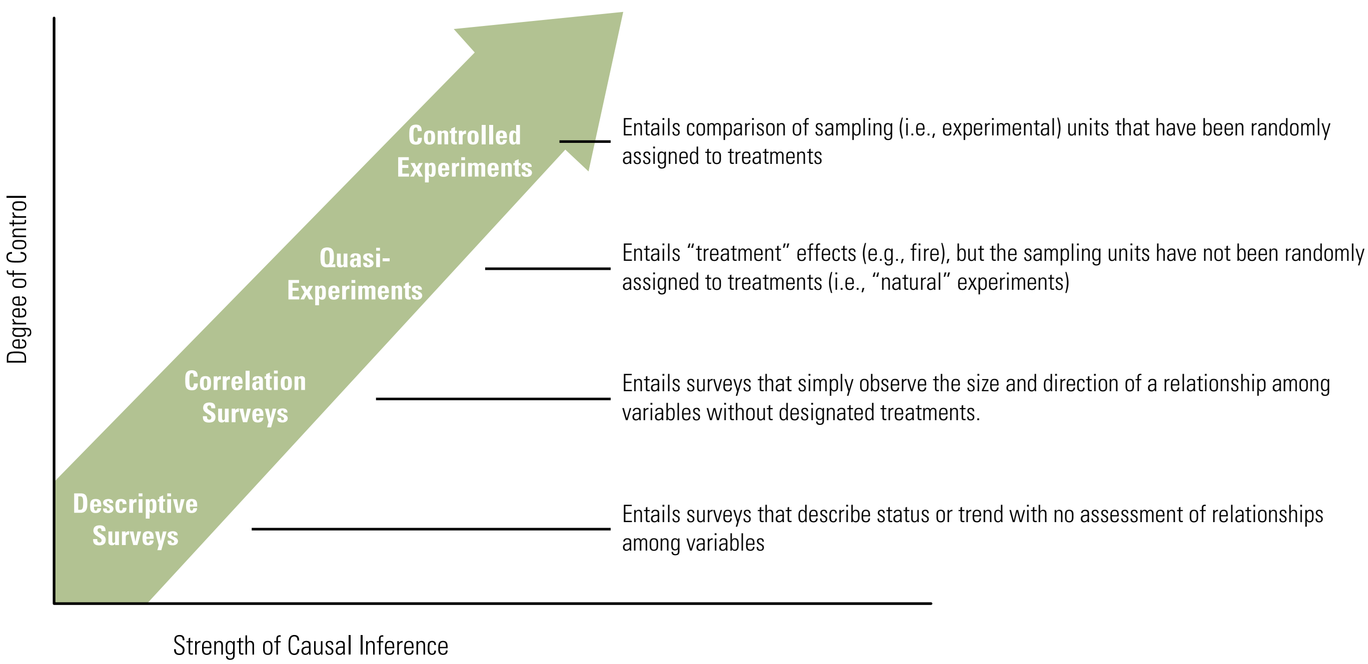


**FIGURE 4.8** Types of surveys or experiments in relation to their degree of control and potential to infer causality (adapted from Schwartz 1998).

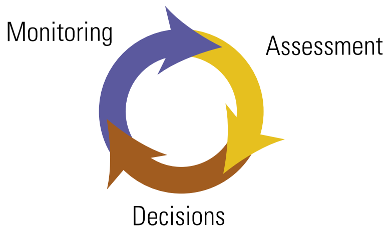


FIGURE 4.9 Conceptual diagram of adaptive management illustrating the iterative cycle of monitoring, assessment and decisions. A science portion of the learning comes from monitoring and assessment and a policy portion comes though incorporating what has been learned into the decision process.

signs themselves in our monitoring. Including auxiliary variables into a monitoring program may increase the precision of parameters of interest by accounting for otherwise unexplained variation (Thompson 2002), particularly when a strong and direct relationship exists between the primary and auxiliary variables (Schwartz 1998). Accounting for such variation may also reduce the risk of misinterpreting results that are artifacts of confounded variables.

*GRYN Example*

For example, landbird monitoring will likely include measurement of vegetation structure as an auxiliary variable(s) with the intent to obtain more precise estimates of the trends in bird abundance by taking into account the confounding effects of vegetation structure. The intent is not to determine if an association exists between vegetation structure and bird abundance or distribution; there have been numerous papers illustrating such an effect. This will improve the precision of the estimates of interest and reduce the chance for misinterpretation of the trends.

## *Design vs. Model-based Inference*

Currently, the objectives that have been created by the GRYN fall primarily under a design-based framework (e.g., Hansen et al. 1983), which uses probability sampling to derive estimates of state variables and/or rates of change. An advantage of this approach is that it minimizes the number of assumptions required to drawn inference, which makes it defensible in cases of litigation and in making controversial public policy decisions (Olsen et al. 1999). However, a design-based approach also tends to be poorly suited for making future predictions. Predictive inferences are generally better suited to a model-based approach (Olsen et al. 1999). A model-based approach enables incorporation of hypothesized relationships, which can better lead to predictive capabilities. However, this advantage comes at a cost of requiring a greater number of simplifying assumptions (Olsen et al. 1999). Additionally, even with reliable parameter estimates, our predictive capabilities will only be as good as the models from which they are derived. However, in the future, a model-based approach may better enable the GRYN to move from a purely descriptive approach to a more scientific (e.g., quasi-experimental) approach to monitoring that can provide advantages for understanding the system and predicting the outcome of management decisions (see also Yoccoz et al. 2001).

Although there can be several advantages in moving toward a model-based approach, it is also premature at this stage of our program development. There are many hypotheses about ecosystem functioning within the GRYN, but in most cases these are not sufficiently conceptualized for an efficient incorporation into a model-based approach. However, as the GRYN program matures, specific hypotheses can be refined and articulated so as to be considered in a model-based context.

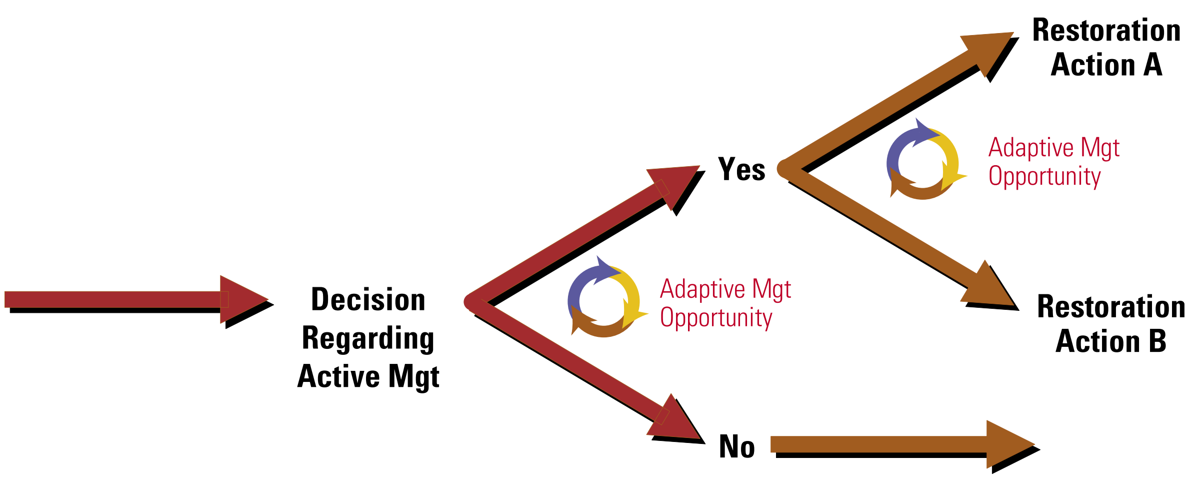


FIGURE 4.10 Conceptual diagram of potential decisions that could benefit from incorporating an adaptive management approach to design.

## Adaptive Management

One goal of monitoring vital signs is to provide the information needed to make better-informed management decisions. Yet, a common mistake of environmental monitoring is a failure to link indicators to decisions (Failing and Gregory 2003). Adaptive management is one tool that could facilitate this linkage between the information derived from monitoring and decisions. Adaptive management is an iterative process of assessment, decision making and monitoring to achieve management goals, whereby learning is facilitated through an experimental approach (e.g., Holling 1978, Walters 1986) (Figure 4.9).

### *GRYN Example*

Whitebark pine is considered a "keystone" species throughout the Greater Yellowstone Ecosystem. It serves a variety of roles ranging from a food source for grizzly bears to having an effect on snow accumulation and distribution. In recent decades whitebark pine stands have been decimated in areas of the Cascades and northern Rocky Mountains due to the introduction of an exotic fungus—white pine blister rust—as well as mountain pine beetles. Our specific monitoring objectives are intended to determine if white pine blister rust is increasing within the Greater Yellowstone Ecosystem, and whether or not the resulting mortality of whitebark pine is sufficient to warrant consideration of management intervention (e.g., active restoration)? Thus, several potential decisions are foreseeable (Figure 4.10). The first decision is whether or not active restoration should be initiated. If it is initiated, there is considerable uncertainty about the effectiveness of alternative management activities for achieving the management objectives of restoration. Both of these decisions have the potential to benefit from an adaptive management approach. If the decision for whether or not to implement active restoration is not universally applied to all areas, then there is the possibility of designing the monitoring to compare management intervention (i.e., active restoration) with an alternative of allowing the process to continue uninterrupted by such intervention. Similarly, if a decision is made to initiate active restoration, then there is an excellent possibility of designing the monitoring to compare alternative restoration practices (e.g., different levels of planting or overstory release, or different types of genetically resistant stock).